

Principle of Detailed Balance and Convergence Assessment of Markov Chain Monte Carlo methods and Simulated Annealing

Ioana A. Cosma and Masoud Asgharian*

July 20, 2008

Abstract

Markov Chain Monte Carlo (MCMC) methods are employed to sample from a given distribution of interest, π , whenever either π does not exist in closed form, or, if it does, no efficient method to simulate an independent sample from it is available. Although a wealth of diagnostic tools for convergence assessment of MCMC methods have been proposed in the last two decades, the search for a dependable and easy to implement tool is ongoing. We present in this article a criterion based on the principle of detailed balance which provides a qualitative assessment of the convergence of a given chain. The criterion is based on the

*Ioana A. Cosma is a doctoral student in the Department of Statistics, University of Oxford, 1 South Parks Road, Oxford, OX1 3TG, United Kingdom (email: cosma@stats.ox.ac.uk); Masoud Asgharian is Associate Professor, Department of Mathematics and Statistics, McGill University, Burnside Hall, 805 Sherbrooke W., Montreal, Quebec, Canada, H3A 2K6 (email: masoud@math.mcgill.ca). This research was partially supported by research grants from NSERC and FQRNT. The authors thank Russell Steele for insightful discussions on the topic.

behaviour of a one-dimensional statistic, whose asymptotic distribution under the assumption of stationarity is derived; our results apply under weak conditions and have the advantage of being completely intuitive. We implement this criterion as a stopping rule for simulated annealing in the problem of finding maximum likelihood estimators for parameters of a 20-component mixture model. We also apply it to the problem of sampling from a 10-dimensional funnel distribution via slice sampling and the Metropolis-Hastings algorithm. Furthermore, based on this convergence criterion we define a measure of efficiency of one algorithm versus another.

KEY WORDS: Metropolis-Hastings; slice sampling; Markov chain Central Limit Theorem; detailed balance; ergodic Markov chain; equilibrium; stationary distribution.

1. INTRODUCTION

Let π be a given distribution such that either π does not exist in closed form or no efficient method to simulate an independent sample from it is available. Suppose that interest lies in the expected value of a random variable $h(X)$, denoted by $\mathbb{E}_\pi(h(X))$, where X has distribution π . Monte Carlo sampling methods (Hammersley and Handscomb 1964) such as rejection sampling, importance sampling or sampling-importance resampling (SIR) approximate the value of $\mathbb{E}_\pi(h(X))$ by sampling from a distribution g that closely resembles π (Smith and Gelfand 1992). Although for low dimensional distributions π it is oftentimes possible to find sampling distributions g that provide estimates to within given accuracy with low computational cost, these sampling methods suffer greatly from the curse of dimensionality.

The need to approximate the value of high dimensional integrals arising in statistical mechanics led to the development of MCMC sampling methods. The first MCMC method, known today as the Metropolis Monte Carlo algorithm, was proposed by

Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller (1953) as a general method for studying the equilibrium properties of systems consisting of many interacting particles. The algorithm simulates the behaviour of the system under equilibrium, and the expected value of a given property is approximated by ergodic averages based on these simulations. In statistical terms, the Metropolis Monte Carlo algorithm constructs an ergodic Markov chain $\{X_t, t = 1, \dots, n\}$ with stationary distribution π , i.e. as the number of iterations n tends to ∞ , the conditional distribution of X_n given the value of X_1 converges to π regardless of the starting distribution g , where X_1 has distribution g (in notation: $X_1 \sim g$).

Hastings (1970) generalized the procedure of proposing the next move X_t given $X_{t-1} = x_{t-1}$. His algorithm, known as the Metropolis-Hastings algorithm, transforms an arbitrary stochastic matrix into a π -reversible one, and only requires that π be known up to a normalizing constant. An equally popular MCMC algorithm is the Gibbs sampler, introduced by Geman and Geman (1984) with an application to image restoration. This algorithm proposes the next move by sampling from the full conditional distributions and, unlike the Metropolis-Hastings algorithm, accepts each proposal with probability 1. Two well-known variants on Gibbs sampling are the data-augmentation algorithm of Tanner and Wong (1987) and the substitution sampling algorithm of Gelfand and Smith (1990).

The goal of MCMC methods is to produce an approximate i.i.d. sample $\{X_{K+1}, X_{K+2}, \dots, X_{K+n}\}$ from π , where $K, n > 1$, and K is known as the number of ‘burn-in’ iterations to be removed from the beginning of the chain. Analysing the output of an MCMC method consists of assessing convergence to sampling from π , convergence to i.i.d. sampling, and convergence of empirical averages of the form $\frac{1}{n} \sum_{i=1}^n h(X_{K+i})$ to $\mathbb{E}_\pi(h(X)) = \int h(x)\pi(x)dx$ as $n \rightarrow \infty$. Robert and Casella (2004) argue that while convergence to π is not of major concern since it can only be achieved

asymptotically, the issues of convergence to i.i.d. sampling and of convergence of empirical averages are strongly interrelated and depend on the mixing speed of the chain. By definition, a chain whose elements converge rapidly to weakly correlated draws from the stationary distribution is said to possess good mixing speed. Therefore, the mixing speed of a chain is determined by the degree to which the chain escapes the influence of the starting distribution and by the extent to which it explores the high density regions of the support of π .

Recent research in MCMC methodology has focused on developing, on one hand, samplers that escape quickly the attraction of the starting distribution as well as that of local modes, and, on the other hand, convergence assessment criteria for analysing the mixing speed of a given chain. A recent sampling algorithm which exploits the idea of jumping between states of similar energy to facilitate efficient sampling is the equi-energy sampler of Kou *et al.*(2006). Robert (1995,1998), Cowles and Carlin (1996), and Brooks and Roberts (1998) present a comprehensive review of the practical implementation of convergence criteria and the mathematics underlying them. Liu (2001), Neal (1993), Brooks (1998), and Kass, Carlin, Gelman, and Neal (1998) offer an in-depth introduction to MCMC methodology and its applications, as well as discussions on the issues surrounding it.

The common view among researchers and practitioners is that developing a good sampler or a reliable convergence criterion is problem-specific. A sampler with good mixing speed when sampling from a relatively smooth, low-dimensional distribution might become trapped in a well of low probability when sampling from a distribution having many local modes. Similarly, a convergence criterion which proves reliable for analysing a given MCMC output might incorrectly assess the convergence of a chain that has only explored a subset of the entire support space. Our interest lies in convergence assessment, in particular, in identifying lack of convergence. We define

a one-dimensional statistic and derive an intuitive criterion based on the principle of detailed balance that provides a qualitative assessment on the convergence of a given MCMC chain.

In Section 2 we recall basic notions and results from the theory of Markov chains, which we subsequently use in Section 3 to derive the asymptotic distribution of our proposed statistic under the assumption of stationarity. In the same section, we discuss two possible implementations of our criterion, one using the asymptotic distribution, the other experimental as a qualitative tool. Section 4 discusses two applications: one as a stopping rule for simulated annealing, an algorithm for function maximization applied to the problem of finding maximum likelihood estimators (Azencott 1992), the second as a graphical tool for comparing the performances of Metropolis-Hastings versus slice sampling for the problem of sampling from a 10-dimensional funnel distribution. All computations were performed using code written in C++. We conclude in Section 5 with general remarks, comparisons, and criticisms.

2. PRELIMINARIES

Let $X = \{X_t, t = 1, 2, \dots\}$ be a Markov chain with state space S and transition probability matrix $P = (p_{ij})$. We refer the reader to Medhi (1994), Norris (1997), and Jones (2004) for details and proofs. For the purpose of the convergence criterion we present in this article, we restrict our attention to finite Markov chains.

Let $p_{ij}^{(n)}$ be the transition probability from state i to state j in n steps. The Ergodic Theorem states that if X is irreducible and aperiodic, then the limits $\pi_j := \lim_{n \rightarrow \infty} p_{ij}^{(n)}$ exist and are independent of the initial state i for all $i, j \in S$ and $(\pi_j, j \in S)$ is the stationary distribution of X . The chain X is called ergodic.

Definition 1 (*Principle of detailed balance*) *Transition probability matrix P and probability distribution π are said to be in detailed balance, or, equivalently, the principle of detailed balance is said to hold, if $\pi_i p_{ij} = \pi_j p_{ji} \forall i, j \in S$.*

Definition 2 *A Markov chain X with irreducible transition probability matrix P and initial distribution g , i.e. $X_1 \sim g$, is reversible if, for all $N \geq 2$, the chain $\{X_N, X_{N-1}, \dots, X_2, X_1\}$ is a Markov chain with transition probability matrix P and initial distribution g .*

Norris (1997) proves that if X is irreducible, then it is reversible if and only if P and g are in detailed balance, where g is the initial distribution of X . The following definitions are needed to introduce the Markov chain Central Limit Theorem (Jones 2004).

Definition 3 *Let $M(i)$ be a nonnegative function and $\gamma(n)$ a nonnegative decreasing function on the positive integers such that*

$$\|P^n(i, \cdot) - \pi(\cdot)\| \leq M(i)\gamma(n). \quad (1)$$

Let X be a Markov chain on state space S with transition probability P and stationary distribution π . If (1) holds for all $i \in S$ with $\gamma(n) = t^n$ for some $t < 1$, then X is geometrically ergodic. If, moreover, M is bounded, then X is uniformly ergodic. If (1) holds for all $i \in S$ with $\gamma(n) = n^{-m}$ for some $m \geq 0$, then X is polynomially ergodic of order m .

Theorem 1 *The **Central Limit Theorem** (finite state space) Let X be an ergodic Markov chain on state space S with stationary distribution π . Let $h : S \rightarrow \mathbb{R}$ be a Borel function. Assume that one of the following conditions holds:*

- 1. X is polynomially ergodic of order $m > 1$, $E_\pi M < \infty$ and there exists $B < \infty$ such that $|h(X)| < B$ almost surely;*
- 2. X is polynomially ergodic of order m , $E_\pi M < \infty$ and $E_\pi(|h(X)|^{2+\delta}) < \infty$ where $m\delta > 2 + \delta$;*

3. X is geometrically ergodic and $E_\pi(|h(X)|^{2+\delta}) < \infty$ for some $\delta > 0$;
4. X is geometrically ergodic and $E_\pi(h^2(X)[\log^+ |h(X)|]) < \infty$;
5. X is geometrically ergodic, satisfies detailed balance and $E_\pi h^2(X) < \infty$;
6. X is uniformly ergodic and $E_\pi(h^2(X)) < \infty$.

Then for any initial distribution,

$$\sqrt{n}(\bar{h}_n - E_\pi(h(X))) \xrightarrow{\mathcal{D}} \text{Normal}(0, \sigma_h^2) \text{ as } n \rightarrow \infty,$$

where $\bar{h}_n = \frac{1}{n} \sum_{i=1}^n h(X_i)$ and $\sigma_h^2 = \text{var}_\pi(h(X_1)) + 2 \sum_{i=2}^\infty \text{cov}_\pi(h(X_1), h(X_i)) < \infty$.

3. DETAILED BALANCE AND CONVERGENCE DIAGNOSTICS

Let $\pi = (\pi_i, i \in S)$ be a discrete distribution with finite state space S , $m = |S|$. Let $\{X_t, t = 1, \dots, n\}$ be an irreducible, aperiodic Markov chain with transition probability matrix $P = (p_{ij})$ and stationary distribution π . We say that a chain has reached equilibrium by step t if $P^t(i, j) = \pi_j, \forall i, j \in S$ and $\exists i, j \in S$ such that $P^{t-1}(i, j) \neq \pi_j$. Our convergence assessment criterion is based on the principle of detailed balance from statistical mechanics (Chandler 1987). Statistical mechanics is concerned with the study of physical properties of systems consisting of very large number of particles, for example liquids or gases, as these systems approach the equilibrium state, i.e. a uniform, time-independent state. In these terms, the principle of detailed balance states that a physical system in equilibrium satisfies

$$\frac{\pi_i}{\pi_j} = \frac{p_{ji}}{p_{ij}} = \exp\left(-\frac{E_i - E_j}{kT}\right), \forall i, j \in S,$$

where E_i is the energy of the system in state i , k is Boltzmann's constant, T is the temperature, and π_i and p_{ij} have the usual interpretation.

We assume that the Markov chain $\{X_t, t = 1, \dots, n\}$ is constructed to satisfy detailed balance. This is oftentimes the case since the principle of detailed balance implies that π is the stationary distribution of the chain, and it is easier to check the former than the latter, see for example the discussions on the Metropolis-Hastings (Hastings 1970) and slice sampling algorithms (Neal 2003). We introduce the notion of an energy function $E_i \propto -\log(\pi_i)$, $\forall i \in S$. When implementing simulated annealing, the stationary distribution at temperature T_k is π^{1/T_k} , so the energy function becomes $E_i = -\log(\pi_i)/T_k$, where $\{T_k, k = 1, 2, \dots\}$ is a sequence of decreasing temperatures. Therefore, the equilibrium probability of being in state i equals $\pi_i = \frac{1}{Z} \exp(-E_i)$, where the normalizing constant is defined as $Z := \sum_{i \in S} \exp(-E_i)$. Define the following approximation to π_i based on a Markov chain of n iterations

$$\hat{\pi}_i = \frac{1}{n} \sum_{j=1}^n \mathbb{I}(X_j = i), \quad \forall i \in S.$$

The idea of working with indicator functions is similar to that of Raftery and Lewis (1992) who develop a convergence assessment method based on the sequence $\{\mathbb{I}(X_t \leq i), t = 1, \dots\}$, for fixed $i \in S$. We point out that, for fixed $i \in S$, the sequence $\{\mathbb{I}(X_t = i), t = 1, \dots\}$ forms a Markov chain, whereas the sequence defined by Raftery and Lewis does not. Brooks *et al.* (2003) use a similar approach of estimating the stationary distribution by the empirical distribution function obtained from the MCMC output; they derive nonparametric convergence assessment criteria for MCMC model selection by monitoring the distance, as the number of simulations increases, between the empirical mass functions obtained from multiple independent chains.

Our criterion assesses the convergence of the chain by comparing the behaviour of the functions $f_i = \hat{\pi}_i / \exp(-E_i)$, $i \in S$, to their average $\bar{f} = \frac{1}{m} \sum_{j \in S} f_j$, via the statistic $V_n := \frac{n}{m} \sum_{i \in S} (f_i - \bar{f})^2$.

3.1 Theoretical approach

We proceed to derive the distribution of the statistic V_n under the hypothesis that the chain has reached stationarity, i.e. that $X_i \sim \pi$, $\forall i = 1, \dots, n$.

$$V_n = \frac{n}{m} \sum_{i \in S} \left\{ f_i - \frac{1}{m} \sum_{j \in S} f_j \right\}^2 = \frac{n}{m} \sum_{i \in S} \left\{ f_i - \frac{1}{m} f_i - \frac{1}{m} \sum_{\substack{j \in S \\ j \neq i}} f_j \right\}^2 = \frac{n}{m} \sum_{i \in S} \{\mathbf{a}_i' \mathbf{f}\}^2,$$

where $\mathbf{f} = (f_i, i \in S)'$ and $\mathbf{a}_i = (-\frac{1}{m}, \dots, -\frac{1}{m}, 1 - \frac{1}{m}, -\frac{1}{m}, \dots, -\frac{1}{m})'$ is an m -dimensional column vector with i th entry equal to $1 - \frac{1}{m}$ and the remaining entries equal to $-\frac{1}{m}$. Define the following $(m \times m)$ dimensional matrix

$$\mathbf{A} = \begin{pmatrix} \mathbf{a}_1' \\ \mathbf{a}_2' \\ \vdots \\ \vdots \\ \mathbf{a}_m' \end{pmatrix} = \begin{pmatrix} 1 - \frac{1}{m} & -\frac{1}{m} & -\frac{1}{m} & \cdots & -\frac{1}{m} \\ -\frac{1}{m} & 1 - \frac{1}{m} & -\frac{1}{m} & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ -\frac{1}{m} & \cdots & \cdots & 1 - \frac{1}{m} & -\frac{1}{m} \\ -\frac{1}{m} & \cdots & \cdots & -\frac{1}{m} & 1 - \frac{1}{m} \end{pmatrix},$$

so $V_n = \frac{n}{m} \{\mathbf{A}\mathbf{f}\}' \{\mathbf{A}\mathbf{f}\}$.

First, we observe that $\forall i \in S$,

$$(f_j - \mathbb{E}_\pi f_j, j \in S) \mathbf{a}_i' = \left(1 - \frac{1}{m}\right) [f_i - \mathbb{E}_\pi f_i] - \frac{1}{m} \sum_{\substack{j \in S \\ j \neq i}} (f_j - \mathbb{E}_\pi f_j) = f_i - \bar{f}, \quad (2)$$

since $\mathbb{E}_\pi f_j = \frac{1}{Z}$, $\forall j \in S$. Second, we notice that

$$f_i - \mathbb{E}_\pi f_i = \frac{\hat{\pi}_i}{e^{-E_i}} - \frac{1}{Z} = \frac{\hat{\pi}_i - \pi_i}{Z\pi_i}, \quad \forall i \in S. \quad (3)$$

Define $W_{i,n} := \sqrt{n}(\hat{\pi}_i - \pi_i)$, $\forall i \in S$, and the m -dimensional column vector $W_n := (W_{i,n}, i \in S)'$. From (2) and (3), we obtain that $V_n = \{\mathbf{C}W_n\}' \{\mathbf{C}W_n\}$, where

$$\mathbf{C} = \mathbf{A} \begin{pmatrix} \frac{1}{\sqrt{mZ\pi_1}} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sqrt{mZ\pi_2}} & 0 & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \frac{1}{\sqrt{mZ\pi_m}} \end{pmatrix} = \begin{pmatrix} \frac{m-1}{m^{3/2}e^{-E_1}} & 0 & \cdots & 0 \\ 0 & \ddots & 0 & \vdots \\ \vdots & 0 & \ddots & \vdots \\ 0 & \cdots & 0 & \frac{m-1}{m^{3/2}e^{-E_m}} \end{pmatrix}$$

The following result presents the asymptotic distribution of the statistic V_n under the assumption of stationarity.

Theorem 2 *Under the conditions of Theorem 1, $\mathbf{C}W_n \xrightarrow{\mathcal{D}} \text{Normal}(\mathbf{0}, \mathbf{C}\Sigma\mathbf{C}')$ and $V_n \xrightarrow{\mathcal{D}} \sum_{i=1}^k \lambda_i Z_i^2$ as $n \rightarrow \infty$, where $\lambda_1, \dots, \lambda_k$ are the characteristic roots of $\mathbf{C}\Sigma\mathbf{C}'$ and Z_1, \dots, Z_k are i.i.d. $\text{Normal}(0, 1)$ random variables.*

proof: We begin by pointing out that irreducible and aperiodic Markov chains on finite state spaces are uniformly ergodic (Roberts and Rosenthal 2004), so condition (6) of Theorem 1 is satisfied. It follows that for every $i \in S$,

$$W_{i,n} = \sqrt{n}(\hat{\pi}_i - \pi_i) = \sqrt{n} \left\{ \frac{1}{n} \sum_{j=1}^n \mathbb{I}(X_j = i) - \mathbb{E}_\pi(\mathbb{I}(X_1 = i)) \right\} \xrightarrow{\mathcal{D}} \text{Normal}(0, \sigma_i^2)$$

as $n \rightarrow \infty$, where

$$\sigma_i^2 = \pi_i(1 - \pi_i) + 2 \sum_{j=2}^{\infty} \left[P\{\mathbb{I}(X_j = i) = 1 | \mathbb{I}(X_1 = i) = 1\} \pi_i - \pi_i^2 \right] < \infty.$$

By the Cramér-Wold Device (Billingsley 1968, Varadarajan 1958), it follows that $W_n \xrightarrow{\mathcal{D}} \text{Normal}(\mathbf{0}, \Sigma)$ as $n \rightarrow \infty$, where $\mathbf{0}$ is an m -dimensional column vector of zeros and Σ is an $(m \times m)$ variance-covariance matrix whose entries are given

$$\begin{aligned} \Sigma(i, i) &= \sigma_i^2 \\ \Sigma(i, j) &= \lim_{n \rightarrow \infty} \text{cov}_\pi(W_{i,n}, W_{j,n}) = \lim_{n \rightarrow \infty} \left\{ \frac{1}{n} \sum_{k=1}^n \sum_{l=1}^n \text{cov}_\pi(\mathbb{I}(X_k = i), \mathbb{I}(X_l = j)) \right\} \\ &= \lim_{n \rightarrow \infty} \left\{ \frac{1}{n} \sum_{k=1}^n \left[P\{X_k = i, X_k = j\} - \pi_i \pi_j \right] + \frac{1}{n} \sum_{\substack{k, l=1 \\ k < l}}^n \left[P\{X_k = i, X_l = j\} \right. \right. \\ &\quad \left. \left. - \pi_i \pi_j \right] + \frac{1}{n} \sum_{\substack{k, l=1 \\ l < k}}^n \left[P\{X_k = i, X_l = j\} - \pi_i \pi_j \right] \right\} \end{aligned}$$

So, for all $i, j \in S, i \neq j$

$$\begin{aligned}
\Sigma(i, j) &= -\pi_i \pi_j + \lim_{n \rightarrow \infty} \frac{\pi_i}{n} \left\{ \sum_{\substack{k, l=1 \\ k < l}}^n \left[P\{X_l = j | X_k = i\} - \pi_j \right] \right. \\
&\quad \left. + \sum_{\substack{k, l=1 \\ l < k}}^n \left[P\{X_l = j | X_k = i\} - \pi_j \right] \right\} \\
&= -\pi_i \pi_j + 2\pi_i \sum_{k=2}^{\infty} \left[P\{X_k = j | X_1 = i\} - \pi_j \right] < \infty,
\end{aligned}$$

The last equality follows from the fact that if a Markov chain satisfies detailed balance, then it is reversible, i.e. for $k > 1$, $P\{X_k = j | X_1 = i\} = P\{X_1 = j | X_k = i\}$. Finally, the conditions of the Markov chain Central Limit Theorem guarantee that the infinite summation in the last line is finite.

It then follows that $\mathbf{C}W_n \xrightarrow{\mathcal{D}} \text{Normal}(\mathbf{0}, \mathbf{C}\Sigma\mathbf{C}')$ as $n \rightarrow \infty$. Lastly, since $V_n = \{\mathbf{C}W_n\}'\{\mathbf{C}W_n\}$, it follows from Lemma 1 in Chernoff and Lehmann (1953) that $V_n \xrightarrow{\mathcal{D}} \sum_{i=1}^k \lambda_i Z_i^2$ as $n \rightarrow \infty$, where $\lambda_1, \dots, \lambda_k$ are the characteristic roots of $\mathbf{C}\Sigma\mathbf{C}'$ and Z_1, \dots, Z_k are i.i.d. $\text{Normal}(0, 1)$ random variables.

Q.E.D.

Example 1 *Let the Markov chain be generated by the Metropolis-Hastings algorithm with symmetric proposal probability matrix $P = (p_{ij})$. The expressions for $\Sigma(i, i)$ and $\Sigma(i, j)$ can be simplified as follows. Consider the Markov-Bernoulli chain $\{\mathbb{I}(X_j = i), j = 1, \dots, n\}$ for fixed $i \in S$ with transition probability matrix $P_i = \begin{pmatrix} 1-a & a \\ b & 1-b \end{pmatrix}$. It is shown in Medhi (1994, pp. 101-102) that*

$$P_i^{j-1} = \frac{1}{a+b} \begin{pmatrix} b & a \\ b & a \end{pmatrix} + \frac{(1-a-b)^{j-1}}{a+b} \begin{pmatrix} a & -a \\ -b & b \end{pmatrix}, \quad \forall j \geq 2.$$

Now,

$$\begin{aligned} a &= \frac{\sum_{\substack{j \in S \\ j \neq i}} P\{X_1 = j, X_2 = i\}}{1 - P\{X_1 = i\}} = \frac{P\{X_2 = i\} - P\{X_1 = i, X_2 = i\}}{1 - P\{X_1 = i\}} = \frac{\pi_i}{1 - \pi_i} (1 - p_{ii}), \\ b &= 1 - P\{X_2 = i | X_1 = i\} = 1 - p_{ii}. \end{aligned}$$

Then, provided that $\max\{0, 2\pi_i - 1\} < p_{ii} < 1$, $\forall i \in S$,

$$\begin{aligned} \Sigma(i, i) &= \pi_i(1 - \pi_i) + 2 \sum_{j=2}^{\infty} \pi_i(1 - \pi_i) \left(\frac{p_{ii} - \pi_i}{1 - \pi_i} \right)^{j-1} = \frac{\pi_i(1 - \pi_i)(1 + p_{ii} - 2\pi_i)}{1 - p_{ii}}, \\ \Sigma(i, j) &= -\pi_i\pi_j + 2\pi_i \sum_{k=2}^{\infty} \left(P^{k-1}(i, j) - \pi_j \right), \quad \text{for } i \neq j. \end{aligned}$$

3.2 Implementation

Let $\{X_{K+1}, X_{K+2}, \dots, X_{K+n}\}$ be an irreducible and aperiodic Markov chain with finite state space S and stationary distribution π that satisfies detailed balance. A burn-in of K draws are discarded, where K depends on the rate of convergence of the sampling algorithm on π (Brooks 1998). We implement our convergence assessment criterion as a test of hypothesis under the null hypothesis that the chain has reached stationarity by iteration $K + 1$.

For n large enough, $V_n \stackrel{\mathcal{D}}{=} \sum_{i=1}^k \lambda_i Z_i^2$, and we estimate its distribution using Lyapunov's Central Limit Theorem (Loève 1963). Since Z_i is $\text{Normal}(0, 1)$, Z_i^2 is $\chi_{(1)}^2$, so $\mathbb{E}(\lambda_i Z_i^2) = \lambda_i$ and $\text{var}(\lambda_i Z_i^2) = 2\lambda_i^2$, for $i = 1, \dots, k$. Define $Y_i = \lambda_i Z_i^2 - \lambda_i$; $\mathbb{E}(Y_i) = 0$, and $\text{var}(Y_i) = \mathbb{E}(Y_i^2) = 2\lambda_i^2 < \infty$ for $i = 1, \dots, n$. Moreover, $\mathbb{E}(Y_i^3) = -4\lambda_i^3 < \infty$, so $\mathbb{E}|Y_i^3| < \infty$, for $i = 1, \dots, k$. Define $s_k^2 = \sum_{i=1}^k \text{var}(Y_i) = 2 \sum_{i=1}^k \lambda_i^2$. It remains to show that the following condition holds: $\lim_{k \rightarrow \infty} \sum_{i=1}^k \mathbb{E}|Y_i|^3 / s_k^3 = 0$, which is equivalent to showing that

$$\lim_{k \rightarrow \infty} \frac{1}{\left(2 \sum_{i=1}^k \lambda_i^2\right)^{3/2}} \sum_{i=1}^k |\lambda_i|^3 = 0, \quad (4)$$

since $\mathbb{E}|Y_i|^3 = |\lambda_i|^3 \mathbb{E}|Z_i^2 - 1|^3 \approx 8.6916|\lambda_i|^3$, for $i = 1, \dots, k$. So, provided that condition (4) is satisfied, Lyapunov's Central Limit Theorem gives the following result for k and n large enough:

$$V_n \stackrel{\mathcal{D}}{=} \sum_{i=1}^k \lambda_i Z_i^2 \sim \text{Normal}\left(\sum_{i=1}^k \lambda_i, 2 \sum_{i=1}^k \lambda_i^2\right) \text{ approximately.} \quad (5)$$

For the computation of the mean and variance in (5), we resort to the following simplifications

$$\sum_{i=1}^k \lambda_i = \text{trace}(\mathbf{C}\Sigma\mathbf{C}') = \sum_{i=1}^m [\mathbf{C}(i, i)]^2 \Sigma(i, i), \quad (6)$$

$$\sum_{i=1}^k \lambda_i^2 = \left(\sum_{i=1}^k \lambda_i\right)^2 - 2 \sum_{\substack{i,j=1 \\ i < j}}^k \lambda_i \lambda_j, \quad (7)$$

where the first summation in equation (7) is given in (6), and the second is the sum of all the 2-square principal subdeterminants of $\mathbf{C}\Sigma\mathbf{C}'$ (Marcus and Ming 1964, p. 22).

We propose a quantitative assessment of convergence via a test of hypothesis at confidence level $(1 - \alpha)$ using the approximate distribution of V_n given in (5) as follows.

1. Obtain an aperiodic, irreducible Markov chain which satisfies the principle of detailed balance: $\{X_1, X_2, \dots, X_K, \dots, X_{K+n}\}$; discard the first K draws.
2. Compute the statistic $V_n = \frac{n}{m} \sum_{i \in S} (f_i - \bar{f})^2$ from the remaining n draws and the $(1 - \alpha/2)$ quantile $v_{\alpha/2} = \sum_{i=1}^k \lambda_i + z_{\alpha/2} \sqrt{2 \sum_{i=1}^k \lambda_i^2}$.
3. If $V_n < v_{\alpha/2}$, conclude that the chain has reached stationarity at level $(1 - \alpha)$ and stop; else, continue for an additional n iterations and return to step 2, replacing n by $2n$.

In this article we implement the criterion in the form of a qualitative tool for convergence assessment. We iterate the chain and plot the absolute value of the relative difference, $|(V_{(k-1)n} - V_{kn})/V_{(k-1)n}|$, against the number of iterations kn , every n iterations, $k = 1, 2, \dots$. We claim that the chain has reached equilibrium if the relative difference drops below some problem-specific, pre-specified constant $\epsilon > 0$. The value of the constant ϵ is problem-specific because it depends on the distribution of interest π . For a high-dimensional, multi-modal distribution, the value of ϵ might need to be very small in order for this analysis to correctly detect lack of convergence to π , whereas the same value might be too conservative for a one-dimensional, unimodal distribution.

Based on this implementation of the criterion as a qualitative tool, we can define a measure of efficiency of one algorithm against another. Let $\epsilon > 0$ be given. Let $V_n^{(i)}$ be the value of the statistic after n iterations of algorithm i , $i = 1, 2$. Let n_i represent the interval, in iterations, at which the statistic is computed for algorithm i . The measure of efficiency is defined as

$$V_{1,2}^{(\epsilon)} = \frac{\min \left\{ kn_1 : |(V_{(k-1)n_1}^{(1)} - V_{kn_1}^{(1)})/V_{(k-1)n_1}^{(1)}| < \epsilon \right\}}{\min \left\{ kn_2 : |(V_{(k-1)n_2}^{(2)} - V_{kn_2}^{(2)})/V_{(k-1)n_2}^{(2)}| < \epsilon \right\}}.$$

If $V_{1,2}^{(\epsilon)} < 1$, we conclude that algorithm 1 is more efficient than algorithm 2 at level ϵ ; if $V_{1,2}^{(\epsilon)} > 1$, algorithm 2 is more efficient than algorithm 1.

4. APPLICATIONS

4.1 Application 1: multipath changepoint problem

The following application is taken from Asgharian and Wolfson (2001). Let Y_{ij} denote the j th measurement on patient i , where $1 \leq i \leq 100$, $1 \leq j \leq 20$. To each patient there is associated a possibly distinct changepoint τ_i such that measurements $Y_{i1}, Y_{i2}, \dots, Y_{i\tau_i}$ are i.i.d. $\text{Normal}(0, 1)$ random variables and measurements

$Y_{i\tau_i+1}, \dots, Y_{i20}$ are i.i.d. $\text{Normal}(4, 1)$. Let $Z_i = (1, Z_{i1})'$ and $\theta = (\theta_0, \theta_1)'$ denote the covariate vector and the regression coefficient vector, respectively, for patient i , i.e. $Y_{ij} = \theta_0 + \theta_1 Z_{i1}$, $\forall j$. Define parameters $\alpha = \theta_0 + \theta_1$ and $\beta = \theta_0 - \theta_1$. The goal is to find the maximum likelihood estimators (MLE's) of α and β , denoted by $\hat{\alpha}$ and $\hat{\beta}$, respectively. We simulate the data with $\theta_0 = 0$ and $\theta_1 = 1$; the joint log likelihood is bimodal. We let the parameter space be $(-10, 10)^2$, assuming zero mass is placed outside this region, and we discretize the space over a grid of width 0.01.

We apply the algorithm of simulated annealing, introduced by Kirkpatrick, Gelatt, and Vecchi (1983), which performs function optimization through an iterative improvement approach. The algorithm was developed via an analogy with thermodynamics where a substance is melted by a slow annealing process and equilibrium is attained at each temperature until eventually the substance stabilizes at its lowest-energy state. Similarly, in simulated annealing, a global temperature parameter controls the effects of high probability regions under the distribution of interest π . For each T_k in a sequence such that $T_k \rightarrow 0$ as $k \rightarrow \infty$, an MCMC chain with stationary distribution π^{1/T_k} is generated until equilibrium. As the temperature is lowered following a pre-specified schedule, known as the cooling schedule, the effects become more pronounced and the chain stabilizes at its global maximum value or equivalently, lowest energy state (Neal 1993, Brooks and Morgan 1995). Geman and Geman (1984) show that this convergence is guaranteed under a logarithmic cooling schedule, which unfortunately is too slow to be followed in practice.

We implement the algorithm with a geometric cooling schedule $T_{k+1} = T_k/2$, $k = 0, \dots, 5$, and $T_0 = 50$ and zero burn-in. Simulated annealing with a very fast cooling schedule is known as simulated quenching; refer to Catoni (1992) for a discussion on the design of cooling schedules. For $(\alpha, \beta) \in (-10, 10)^2$, the function $f_{\alpha, \beta}^{(k)}$ at temperature T_k is given by $f_{\alpha, \beta}^{(k)} = \hat{\pi}_{\alpha, \beta} / \exp(-E_{(\alpha, \beta)})$.

The aim is to compare the performance of the Metropolis-Hastings sampler in determining the MLE's via simulated annealing with two different methods for proposing the next move. In the first method, we draw uniformly from a cube of length w centered at the current position, where w has the values: $\{12, 7, 4, 2.5, 1.7, 1.2, 0.9, 0.6\}$ for $k = 1, \dots, 8$. These values are set retrospectively to obtain an acceptance rate of approximately 0.4. In the second method, we propose the next move via univariate slice sampling applied to each variable in turn; this algorithm is described briefly in Subsection 4.2. We use the “stepping-out” procedure with an initial interval size of 0.1 at each temperature.

At each temperature, we perform 1000 iterations of the Metropolis-Hastings algorithm, computing the value of V_n every 25 iterations. We obtain the following results: $(\hat{\alpha}^{(1)}, \hat{\beta}^{(1)}) = (1.18, -1.17)$, $E_{(\hat{\alpha}^{(1)}, \hat{\beta}^{(1)})} = 247.645$, and $(\hat{\alpha}^{(2)}, \hat{\beta}^{(2)}) = (1.19, -1.15)$, $E_{(\hat{\alpha}^{(2)}, \hat{\beta}^{(2)})} = 247.645$ for the first and second methods, respectively, which equal the lowest energy value obtained by a systematic grid search. We conclude that both methods correctly identified the MLE's. Figures 1 and 2 display the relative difference in variance; sharp drops indicate that the sampler has jumped to previously unexplored regions of the parameter space, i.e. to points (α, β) for which $\hat{\pi}_{\alpha, \beta}$ is significantly different from $\pi_{\alpha, \beta}$, thus increasing the value of the variance.

We proceed to simulate 50 datasets; for each, we initialize the two chains from the same randomly chosen point. At each temperature level, we compute the value of V_n every 25 iterations until $|(V_{(k-1)n} - V_{kn})/V_{(k-1)n}| < \epsilon$, with $\epsilon = 0.05$. We remark that this value of ϵ is very conservative; ideally, a different value would be employed at each temperature level. We make the following two observations: first, for any given dataset, the lowest energy values reported by the two algorithms differ by at most 0.011 units in magnitude, and, second, the difference between the lowest energy values found by a systematic search and by simulated annealing is at

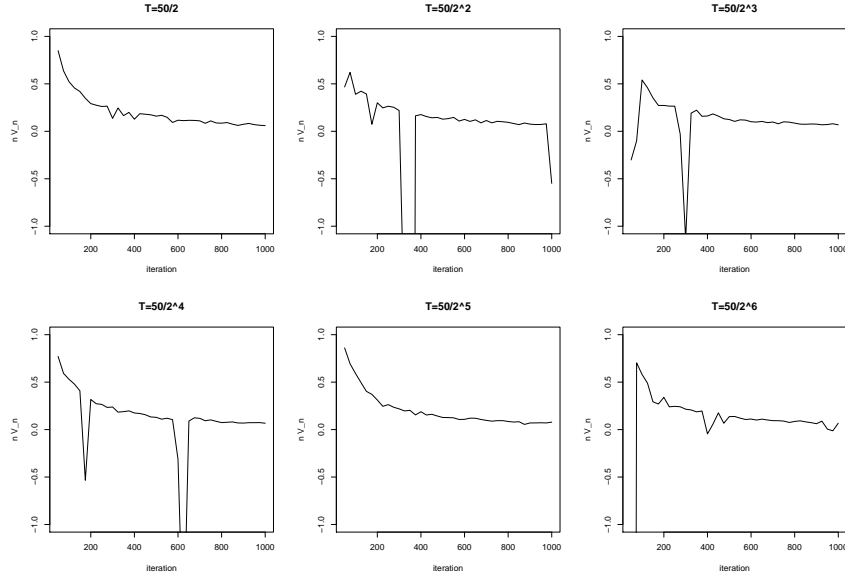


Figure 1: Relative difference in V_n versus n using uniform proposal distributions for application 1. The plots show the decreasing trend of the relative difference in V_n as the number of iterations increases, interrupted by sharp increases in V_n .

most 0.614909. Moreover, we note that the methods required on average 5605 iterations, and 3162 iterations, respectively. Averaged over 50 tests, the measure of efficiency of simulated annealing using Metropolis-Hastings with uniform proposals versus Metropolis-Hastings with slice sampling is approximately 1.77, i.e. MCMC with slice sampling is almost twice as efficient as MCMC with uniform proposals.

4.2 Application 2: 10-dimensional funnel

Neal (2003) illustrates the advantage of slice sampling over Metropolis-Hastings in sampling from a 10-dimensional funnel distribution. Slice sampling is an adaptive MCMC method which proceeds in two alternating steps. Given the current position $X_t = x_t$, it samples a value y uniformly from the interval $(0, \pi(x_t))$. Given y , the next position X_{t+1} is sampled from an appropriately chosen subset of the horizontal “slice” $\{x; \pi(x) > y\}$. Neal (2003) shows that the algorithm produces an ergodic Markov

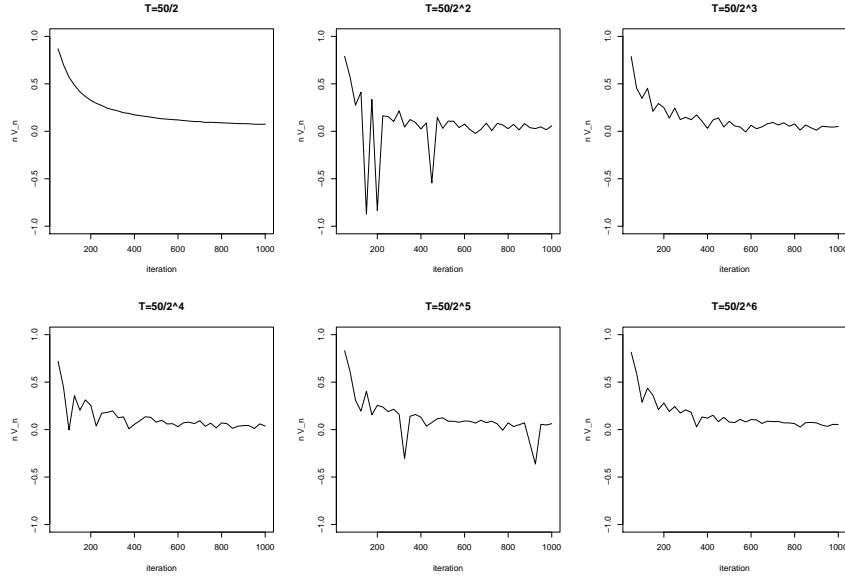


Figure 2: Relative difference in V_n versus n using slice sampling for application 1. The plots show the decreasing trend of the relative difference in V_n as the number of iterations increases; the increases in V_n are more frequent than in Figure 1.

chain with stationary distribution π , and that, moreover, due to its adaptive nature, the algorithm sometimes outperforms Metropolis-Hastings and the Gibbs sampler.

Let X be a $\text{Normal}(0, 9)$ random variable, and let Y_1, \dots, Y_9 be independent Normal random variables, which, conditional on $X = x$, have mean 0 and variance $\exp(x)$. The goal is to obtain an approximate independent sample from the joint distribution of (X, Y_1, \dots, Y_9) . We initialize the chain as follows: $X = 0$ and $Y_i = 1$, for $i = 1, \dots, 9$. For each variable, the parameter space is taken to be $(-30.0, 30.0)$ and it is discretized over a grid of width 0.01.

First, we implement the Metropolis-Hastings algorithm with single-variable updates applied to each variable in sequence; one iteration of the chain consists of 1300 updates. For each variable, the proposal distribution is Normal, centered at the current value, with standard deviation of 1.0, truncated on the interval $(-30.0, 30.0)$.

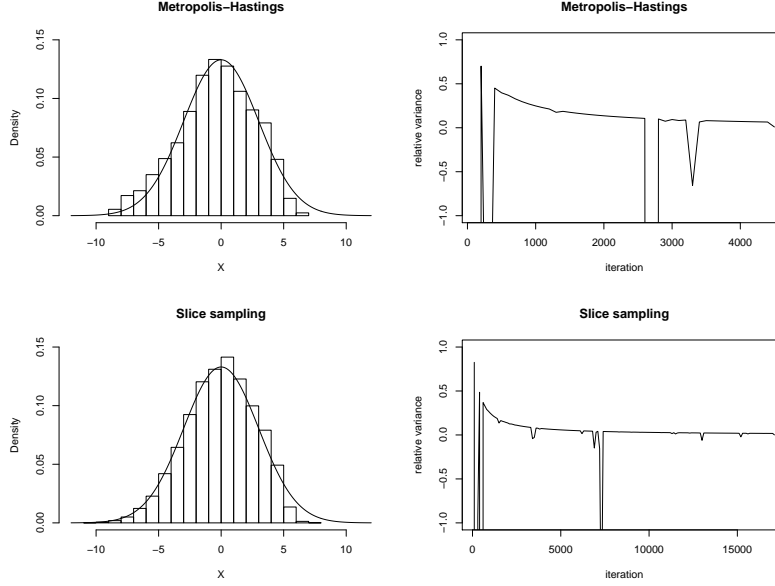


Figure 3: Sampled values and relative difference in V_n in application 2. The left column displays histograms of the sampled values of X superimposed on the Normal(0, 9) density function. The right column displays the relative difference in V_n versus n .

Numbers are rounded to the closest value on the grid. Second, we implement the slice sampling algorithm with single-variable updates; each iteration consists of 120 updates for each variable in sequence. We use the “stepping-out” procedure with an initial interval of size 1. We compute V_n every 100 iterations until the absolute value of the relative difference is below $\epsilon = 0.01$.

The left column of Figure 3 compares the histograms of the sampled values of X with the true probability distribution function; the histograms are based on chains of 4600 and 17200 iterations, respectively. Metropolis-Hastings oversamples negative values of X and undersamples positive ones; slice sampling samples correctly in the left tail of the distribution, but undersamples positive values. The right column displays the behaviour of the relative difference in V_n ; the variance function undergoes sharp increases in value under both sampling methods, but stabilizes towards the end

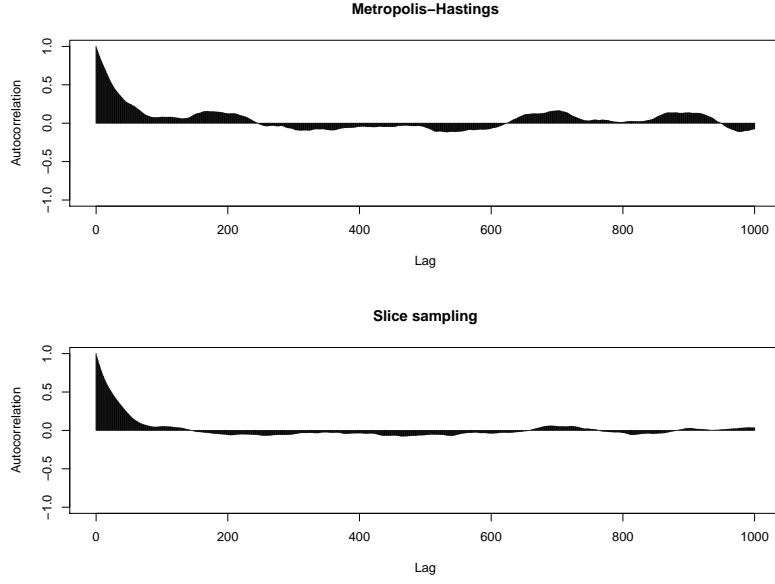


Figure 4: Autocorrelation of X in application 2. Slice sampling has a faster rate of convergence than Metropolis-Hastings evidenced by the smaller autocorrelation.

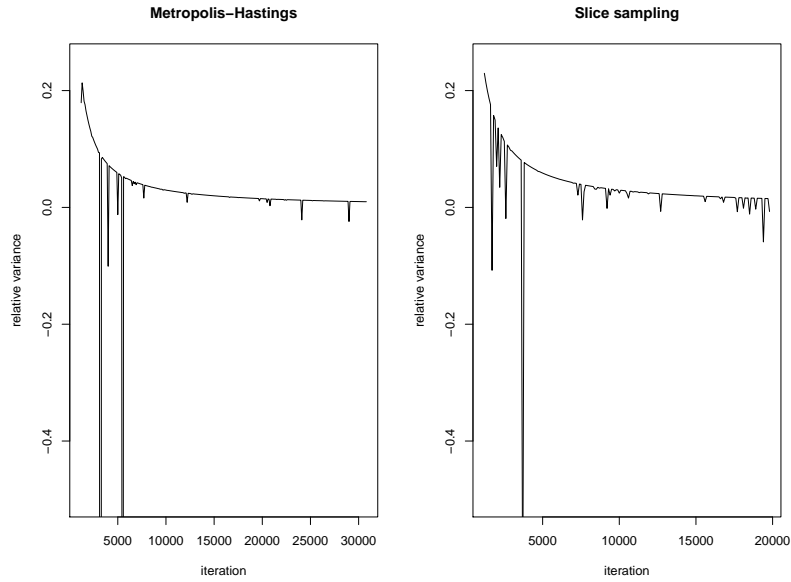


Figure 5: Relative difference in V_n versus n for eleven parallel chains in application 2. The value of V_n under Metropolis-Hastings sampling seems to be more stable than under slice sampling.

of the run. The behaviour of the variance function fails to reflect the incorrect sampling in the tails of the distribution. The plot of the relative difference in variance for the Metropolis-Hastings algorithm indicates that a smaller value of ϵ would be more appropriate for assessing convergence. The plots in Figure 4 show that the autocorrelation obtained by slice sampling remains close to zero after 100 iterations, whereas that obtained by Metropolis-Hastings continues to fluctuate even after 1000 iterations. This indicates that the Metropolis-Hastings algorithm converges more slowly than slice sampling. We compute the Raftery and Lewis (1992) convergence diagnostic using the Coda package in R (<http://www.r-project.org>) obtaining dependence factors of 14 and 18.7 for the Metropolis-Hastings and the slice sampling algorithms, respectively, indicating strong autocorrelation.

Finally, we run eleven parallel chains started from the following quantiles of the marginal distribution of X : $\{0.1, 0.2, 0.3, 0.4, 0.45, 0.5, 0.55, 0.6, 0.7, 0.8, 0.9\}$; we employ the value $\epsilon = 0.01$. We expect the parameter space to be insufficiently explored by both algorithms; however, we are interested in whether this insufficient exploration can be detected from the behaviour of V_n across chains with overdispersed starting points. Pooling the sampled values results in chains of 30800 and 19800 draws, respectively; thus the measure of efficiency of Metropolis-Hastings versus slice sampling is 1.56. Trace plots and histograms indicate that negative values of X are oversampled and positive ones are undersampled by both algorithms. Figure 5 is obtained by pooling the sampled values across the eleven chains; the behaviour of V_n under slice sampling poses signs of concern regarding convergence to stationarity (notice the frequent increases in value from iteration 17500 onwards), whereas the value of V_n under Metropolis-Hastings appears stable towards the end of the run. Therefore the behaviour of V_n under slice sampling across eleven chains with overdispersed starting points indicates lack of convergence to stationarity, whereas the behaviour of V_n

under Metropolis-Hastings, which is known to allow a more restrictive exploration of the support space, gives misleading results.

5. CONCLUSION

The last fifty years have witnessed the development and rise in popularity, in particular in Bayesian statistical inference, of Markov Chain Monte Carlo methods for simulating from complex probability distributions (Smith and Roberts 1993). For a practitioner who has a finite MCMC output, questions arise regarding how reliable the sample is as a representation of π . Although a wealth of convergence diagnostic tools for analysing MCMC output have been proposed over the past decades, their performance, in general, is problem-specific, and developing a dependable, easy to implement tool for convergence assessment continues to be a challenge. This article presents a new convergence assessment method for irreducible, aperiodic Markov chains on discrete spaces obtained by MCMC samplers that satisfy the principle of detailed balance and requirement (4). We introduce a one-dimensional test statistic whose behaviour under the assumption of stationarity is analyzed both theoretically and experimentally, and present a possible implementation of our criterion as a graphical tool for convergence assessment.

In low dimensional problems, the proposed criterion as a qualitative tool assesses convergence satisfactorily; however, in high dimensional problems, the criterion is unreliable for convergence assessment, but can provide useful insight into lack of convergence of the chain to stationarity. In particular, if the variance function experiences sharp increases in value, then it can be concluded that stationarity has not yet been reached; however, if the value of the variance function is stable, then the results are inconclusive. The advantage of our method lies in its attempt to analyse the behaviour of an MCMC chain travelling through a possibly high dimensional space by monitoring the behaviour of a one-dimensional statistic. Lack of convergence to

stationarity is correctly assessed by the behaviour of the statistic to the extent to which the sampler explores freely the underlying space. Particularly in high dimensional problems with irregularly shaped distribution functions, we recommend that the MCMC output be analyzed using different ϵ values, compared across multiple chains, and that several diagnostic tools be employed.

There exist in the literature at least two convergence assessment criteria based on weighting functions that are very similar to our approach. Ritter and Tanner (1992) propose to detect convergence to the full joint distribution by monitoring convergence of the importance weight $w_t = \pi(x)/g_t(x)$, where g_t is the joint distribution of the observations sampled at iteration t . They estimate $g_t(x)$ by $\frac{1}{m} \sum_{i=1}^m p(x|x_{t-1}^{(i)})$, where $x_{t-1}^{(i)}$, $i = 1, \dots, m$ is a sample from g_{t-1} . If the chain has converged, the distribution of the weights w_t , based on multiple replications of the chain, will be degenerate about a constant. Zellner and Min (1995) propose a convergence criterion for the Gibbs sampler in the special case that x can be partitioned into $(x_{(1)}, x_{(2)})$. They define two criteria based on the weight functions $W_1 = p(x_{(1)})p(x_{(2)}|x_{(1)}) - p(x_{(2)})p(x_{(1)}|x_{(2)})$ and $W_2 = [p(x_{(1)})p(x_{(2)}|x_{(1)})] / [p(x_{(2)})p(x_{(1)}|x_{(2)})]$, where $p_{(1)}$ is estimated by $\frac{1}{m} \sum_{i=1}^m p(x_{(1)}|x_{(2)}^i)$, and $x_{(2)}^j$, $j = 1, \dots, m$ is the sequence of draws of $x_{(2)}$ obtained by Gibbs sampling. They compute the value of these weights at many points in the parameter space and argue that if the chain has converged, then the values of W_1 will be close to 0 and those of W_2 close to 1. Zellner and Min use asymptotic results from the stationary time series literature to calculate posterior odds for the hypothesis $H_0 : W_1 = 0$ vs. $H_1 : W_1 \neq 0$ for the k -dimensional case, $k \geq 1$, when the weights are computed at k different points in the parameter space.

The main drawback of these methods is the assumption that the transition probability $p(x|x_{t-1})$, in the method of Ritter and Tanner, and the conditionals $p(x_{(1)}|x_{(2)})$ and $p(x_{(2)}|x_{(1)})$, in the method of Zellner and Min, exist explicitly. Our method, how-

ever, makes no such assumption and estimates π_i , the probability of being in state i , by the empirical distribution function. All three methods have the disadvantage of being computationally expensive; the ergodic averages used to approximate various marginal and conditional probabilities (in our method, $\hat{\pi}_i$) require a large number of summands in order to provide good estimates, so large numbers of iterations, and possibly many replicates of the chain, are needed. Furthermore, since the normalizing constant of π is unknown, the functions f_i and the weights w_t of the criterion of Ritter and Tanner might stabilize around an incorrect value if the sampler has failed to explore all the high density regions of the space. For this reason, we recommend to run multiple replicates of the chain started from different regions of the space. The criterion of Zellner and Min also gives misleading results if the space is poorly explored and the weights are computed at points that come from low density regions. Finally, our criterion has an intuitive graphical representation, very similar to that proposed by Ritter and Tanner, and, whereas the criterion of Zellner and Min uses multivariate weight functions, our criterion is based on a one-dimensional statistic regardless of the dimension of the underlying space, thus offering a dimensionality reduction approach to the problem of convergence assessment in high dimensional spaces.

An interesting alternative to approximating a continuous state space by a discrete grid is to sample the continuous state-space Markov chain and to apply the discretization method developed by Guihenneuc-Jouyaux and Robert (1998). Provided that the continuous chain is Harris-recurrent, the method defines renewal times based on the visiting times to one of m disjoint small sets in the support space. By subsampling the underlying chain at the renewal times, the method builds a homogeneous Markov chain on the finite state space $\{1, \dots, m\}$. Our proposed criterion can then be applied to the finite chain; it would be interesting to explore whether the convergence

assessment extends to the continuous Markov chain.

References

- [1] Asgharian, M. and Wolfson, D. B. (2001) “Modeling covariates in multipath changepoint problems: Modeling and consistency of the MLE,” *The Canadian Journal of Statistics*, 29, 4, 515-528.
- [2] Azencott, R. (ed.) (1992) *Simulated annealing: parallelization techniques*, New York: Wiley.
- [3] Billingsley, P. (1968) *Convergence of Probability Measures*, New York: John Wiley & Sons, Inc.
- [4] Brooks, S. P. (1998) “Markov chain Monte Carlo method and its application,” *The Statistician*, 47, 69-100.
- [5] Brooks, S. P., Giudici, P., and Philippe, A. (2003) “Nonparametric Convergence Assessment for MCMC Model Selection,” *Journal of Computational and Graphical Statistics*, 12, 1, 1-22.
- [6] Brooks, S. P., and Morgan, B. J. T. (1995) “Optimization using simulated annealing,” *The Statistician*, 44, 241-257.
- [7] Brooks, S. P., and Roberts, G. O. (1998) “Convergence assessment techniques for Markov chain Monte Carlo,” *Statistics and Computing*, 8, 319-335.
- [8] Catoni, O. (1992) “Rough large deviation estimates for simulated annealing: application to exponential schedules,” *The Annals of Probability*, 20, 3, 1109-1146.

- [9] Chandler, D. (1987) *Intoduction to Modern Statistical Mechanics*, New York: Oxford University Press.
- [10] Chernoff, H. and Lehmann, E. L. (1953) “The use of maximum likelihood estimates in χ^2 tests for goodness of fit,” *The Annals of Mathematical Statistics*, 25, 3, 579-586.
- [11] Cowles, M. K., and Carlin, B. P. (1996) “Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review,” *Journal of the American Statistical Association*, 91, 883-904.
- [12] Gelfand, A. E., and Smith, A. F. M. (1990) “Sampling-Based Approaches to Calculating Marginal Densities,” *Journal of the American Statistical Association*, 85, 398-409.
- [13] Geman, S., and Geman, D. (1984) “Stochastic Relaxation, Gibbs Distribution, and the Bayesian Restoration of Images,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721-741.
- [14] Guihenneuc-Jouyaux, C., and Robert, C. P. (1998) “Discretization of Continuous Markov Chains and Markov Chain Monte Carlo Convergence Assessment,” *Journal of the American Statistical Association*, 93, 443, 1055-1067.
- [15] Hastings, W. K. (1970) “Monte Carlo sampling methods using Markov chains and their applications,” *Biometrika*, 55, 97-109.
- [16] Hammersley, J. M., and Handscomb, D. C. (1964) *Monte Carlo methods*, London: Methuen.
- [17] Jones, G. (2004) “On the Markov chain central limit theorem”, *Probability Surveys*, 1, 299-320.

- [18] Kass, R. E., Carlin, B. P., Gelman, A., and Neal, R. M. (1998) “Markov Chain Monte Carlo in Practice: A Roundtable Discussion,” *The American Statistician*, 52, 93-100.
- [19] Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P. (1983) “Optimization by Simulated Annealing,” *Science*, 220, 671-680.
- [20] Kou, S. C., Zhou, Q., and Wong, W. H. (2006) “Equi-energy Sampler with Applications in Statistical Inference and Statistical Mechanics,” *The Annals of Statistics*, 34, 4, 1581-1619.
- [21] Loève, M. (1963) *Probability Theory*, Toronto: D. Van Nostrand Company (Canada), Ltd.
- [22] Liu, J. S. (2001) *Monte Carlo strategies in scientific computing*, New York: Springer.
- [23] Marcus, M., and Ming, H. (1964) *A survey of matrix theory and matrix inequalities*, New York: Dover Publications, Inc.
- [24] Medhi, J. (1994) *Stochastic Processes*, New Delhi: New Age International (P) Ltd., second edition.
- [25] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953) “Equation of State Calculations by Fast Computing Machines,” *The Journal of Chemical Physics*, 21, 1087-1092.
- [26] Neal, R. M. (1993) *Probabilistic Inference Using Markov Chain Monte Carlo Methods*, Technical Report CRG-TR-93-1, Department of Computer Science, University of Toronto.

- [27] ———(2003) “Slice Sampling,” *The Annals of Statistics*, 31, 3, 705-767 (with discussion and a rejoinder by the author).
- [28] Norris, J. R. (1997) *Markov Chains*, New York: Cambridge University Press.
- [29] Raftery, A. E., and Lewis, S. (1992) “How Many Iterations in the Gibbs Sampler?”, in *Bayesian Statistics 4*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, Oxford, U.K.: Oxford University Press, 763-773.
- [30] Ritter, C., and Tanner, M. A. (1992) “Facilitating the Gibbs Sampler: The Gibbs Stopper and the Griddy-Gibbs Sampler,” *Journal of the American Statistical Association*, 87, 861-868.
- [31] Robert, C. P. (1995) “Convergence Control Methods for Markov Chain Monte Carlo Algorithms,” *Statistical Science*, 10, 3, 231-253.
- [32] ———(ed.) (1998) *Discretization and MCMC Convergence Assessment*, Lecture Notes in Statistics, 135, New York: Springer.
- [33] Robert, C. P., and Casella, G. (2004) *Monte Carlo Statistical Methods*, New York: Springer-Verlag, second edition.
- [34] Roberts, G. O., and Rosenthal, J. S. (2004) “General state space Markov chains and MCMC algorithms,” *Probability Surveys*, 1, 20-71.
- [35] Smith, A. F. M., and Gelfand, A. E. (1992) “Bayesian Statistics Without Tears: A Sampling-Resampling Perspective,” *The American Statistician*, 26, 84-88.
- [36] Smith, A. F. M., and Roberts, G. O. (1993) “Bayesian Computation via the Gibbs Sampler and Related Markov Chain Monte Carlo Methods,” *Journal of the Royal Statistical Society, Ser. B*, 55, 3-23.

- [37] Tanner, M. A., and Wong, W. H. (1987) “The Calculation of Posterior Distributions by Data Augmentation,” *Journal of the American Statistical Association*, 82, 528-540.
- [38] Varadarajan, V. S. (1958) “A Useful Convergence Theorem”, *Sankhya*, 20, 221-222.
- [39] Zellner, A., and Min, C. (1995) “Gibbs Sampler Convergence Criteria,” *Journal of the American Statistical Association*, 90, 921-927.